Computational Brain & Behavior
https://doi.org/10.1007/s42113-018-0008-2

An Instance Theory of Semantic Memory

Randall K. Jamieson[1] · Johnathan E. Avery[2] · Brendan T. Johns[3] · Michael N. Jones[2]

Abstract
Distributional semantic models (DSMs) specify learning mechanisms with which humans construct a deep representation of word meaning from statistical regularities in language. Despite their remarkable success at fitting human semantic data, virtually all DSMs may be classified as prototype models in that they try to construct a single representation for a word's meaning aggregated across contexts. This prototype representation conflates multiple meanings and senses of words into a center of tendency, often losing the subordinate senses of a word in favor of more frequent ones. We present an alternative instance-based DSM based on the classic MINERVA 2 multiple-trace model of episodic memory. The model stores a representation of each language instance in a corpus, and a word's meaning is constructed on-the-fly when presented with a retrieval cue. Across two experiments with homonyms in both an artificial and natural language corpus, we show how the instance-based model can naturally account for the subordinate meanings of words in appropriate context due to nonlinear activation over stored instances, but classic prototype DSMs cannot. The instance-based account suggests that meaning may not be something that is created during learning or stored per se, but may rather be an artifact of retrieval from an episodic memory store.

Keywords Semantic memory · Computational model · Exemplar-based model · Episodic memory

Introduction

Distributional semantic models (DSMs) such as BEAGLE, HAL, LSA, and Word2Vec represent a major advance in the field of semantic memory (Jones and Mewhort 2007; Lund and Burgess 1996; Landauer and Dumais 1997; Mikolov et al. 2013). DSMs attempt to explain how humans transform first-order statistical experience with language into deep knowledge representations of word meaning. They predict a broad class of behavioral phenomena and have been used to develop cognitive technologies for a number of applied problems (e.g., Aujla et al. 2018; Bedi et al. 2015; Foltz et al. 1999; Johns et al. 2013; Rubin et al. 2016a, b). The mechanisms posited by DSMs to transform episodic experience to semantic representations vary widely, ranging from simple co-occurrence counting to error-driven reinforcement learning (see Jones et al. 2006, for a review).

✉ Michael N. Jones
jonesmn@indiana.edu

[1] University of Manitoba, Winnipeg, Canada

[2] Department of Psychological and Brain Sciences, Indiana University, Bloomington, USA

[3] University at Buffalo, Buffalo, USA

Virtually, all DSMs share one commonality: *They are prototype models*. That is, the representation of a word is collapsed into a single averaged representation. This shared characteristic may represent a significant architectural flaw in DSMs, leading the field to assume that semantic abstraction is a learning mechanism rather than a retrieval mechanism.

All current spatial DSMs use the co-occurrence regularities of words across contexts in language to build a single vector representation that best represents the word's aggregate meaning, formalizing the classic notion that, "you shall know a word by the company it keeps" (Firth 1957). However, the notion of building a single prototypical center of tendency disagrees with the current state-of-the-art in related fields of cognition, such as categorization and episodic memory. The categorization literature, for example, has largely converged on the superiority of exemplar-based theories over prototype theories because prototype theories cannot explain human behavior when dealing with category structures that have nonlinearly separated structure, such as in classic XOR. Even if linear category structures are used that should be optimal for prototype models, exemplar models produce a superior quantitative prediction of human data (e.g., Stanton et al. 2002).

Jones (2017) has recently suggested that current "abstraction-at-learning" DSMs suffer from the same issues as prototype theories in categorization, a problem that arises from collapsing the many contexts in which a word occurs to a

Published online: 16 July 2018

single best-fitting representation. Doing so discards idiosyncratic regularities that are important to word meaning. Here, we argue that homonyms present an ideal method to define and evaluate the potential shortcoming.

It has long been known that spatial DSMs collapse the multiple senses of a homonym into a single representation, averaging over different and distinct context patterns to derive a center of tendency that represents the word's average meaning. For example, a homonym such as *bark* is positioned in semantic space as a frequency weighted average of its distinct senses (e.g., the sound a dog makes versus the outer shell of a tree). This is a graded problem across the continuous modulations in meaning that are inherent to polysemous words as well. Jones suggests that this behavior is not a simple problem that can be handily solved with a patch in abstraction-at-learning DSMs, but rather indicates that the models may be fundamentally wrong in how they conceive of semantics. In short, although they can represent the average meaning of a word, they do so at the cost of being able to represent meanings that differ from the average (e.g., rare senses of homonyms and polysemes), a behavior that is rapid and natural to humans.

In this paper, we develop and test an alternate notion of abstraction in a DSM. Building on established and successful instance-based episodic memory models, like Hintzman's (1984, 1986, 1988) MINERVA 2 model of memory, we posit that semantic abstraction may be a consequence of retrieval from episodic memory rather than a learning mechanism. To make the case, we developed and tested an instance-based theory of semantics that stores word contexts as multiple traces in episodic memory and derives meaning on-the-fly at retrieval.

In contrast to abstraction-at-learning DSMs, an instance-based model can produce nonlinear activation of stored instances, which allows it to access the subordinate sense of a word when provided the appropriate cue (e.g., *bank* as in the sense of turning a plane rather than *bank* in the sense of financial instructions). The model is able to account for traditional phenomena that have been used as support for DSMs. But, it can also explain patterns of responses to subordinate meanings of a word in context that are difficult to account for with traditional DSMs, and to do so without the requirement for an explicit store for semantic memory per se. We demonstrate the model using an artificial language corpus and, then, using a natural language corpus.

"Abstraction-at-Learning" DSMs

The vast majority of spatial DSMs assume that semantic abstraction is a learning mechanism, and the task of the model is to collapse across idiosyncratic linguistic episodes to derive a stable prototypical representation of the word's meaning. The classic example is seen in Landauer and Dumais' (1997) latent semantic analysis (LSA). LSA represents the first-order "episodic" contexts in a word-by-context frequency matrix. In this initial matrix, words are similar only if they frequently co-occur in contexts (i.e., in the same documents across the corpus). LSA then applies singular-value decomposition, a technique from linear algebra, to this episodic matrix and retains the 300 dimensions with the highest eigenvalues. The resulting word vectors emphasize second-order relationships that were latent in the original episodic matrix so that, in the reduced space, words will be similar if they appear in similar contexts, even if they never co-occur directly in the corpus (e.g., synonyms and category coordinates). In summary, LSA uses linear algebra to collapse a word's episodic contexts into a single point in a high-dimensional semantic space.

A similar pattern can be seen across contemporary DSMs. Jones and Mewhort's (2007) BEAGLE model accumulates random vectors across episodic contexts to produce a distributed semantic representation for each word, so that a word's semantic representation is the average of the other words that it has co-occurred with. Before learning, each word that can be encountered in the corpus is assigned a random Gaussian vector to represent its physical characteristics, such as orthography or phonology. This environmental vector is static and remains the same each time the word is encountered. For each studied context, a word's memory vector is encoded as the sum of the environmental vectors for the other words that it co-occurred with. Across many contexts, a word's memory vector becomes a distributed pattern of features that reflects its history of co-occurrence with other words. The final semantic vector for a word is a linear average that tends to emphasize higher-order semantic relationships of words that co-occurred with the same words (e.g., synonyms).

The newest additions to the DSM family are predictive neural embedding models, which use a connectionist architecture and error-driven backpropagation to learn a distributed vector pattern for a word's meaning. The current frontrunner is the Word2Vec model of Mikolov et al. (2013). Word2Vec is a three-layer connectionist network with localist input and output layers (i.e., with one node for each word in the vocabulary), fully connected via a hidden layer of 300 nodes. Each context, a target output word is predicted by using the other co-occurring words as input. The error signal is backpropagated through the network to increase the probability of the network predicting the correct output word given the same input words in future epochs. After many learning epochs, the network settles, and the matrix of hidden-to-output connections is exported as a semantic representation. In this matrix, words are similar if they are predicted by similar contexts. Hence, Word2Vec produces a similar outcome to both LSA and BEAGLE by collapsing a word's episodic contexts into a single reduced representation of meaning. All three models accomplish the same task of producing a prototype for a word's meaning, albeit by different learning mechanisms. To be clear, Word2Vec is a multilayer network; hence, it can

predict the same output word (e.g., *bark*) given very different input patterns (e.g., *bark* in the tree sense versus *bark* in the dog sense). But the final representation for a word is the complete pattern of weights—the prototype. Two words are typically compared via the cosine of their respective vectors in all models, and so a homonym such as *bark* will be pulled between its two senses just as it is in LSA or BEAGLE.

Griffiths et al. (2007; see also Griffiths et al. 2005) have suggested that homonyms present a core challenge to spatial DSMs that the models cannot adequately explain, arguing instead for probabilistic topic models. In addition, homonyms and polysemes are hardly rare in language: The majority of words in English have multiple senses, and the frequency distribution of senses for a word is skewed, loosely conforming to a Zipfian distribution. Prototype DSMs lose the tail when collapsing to a prototype, but humans can regularly comprehend the multiple less frequent meanings that are averaged out in DSMs. Hence, DSMs have great difficulty with the subordinate senses of homonyms (e.g., the river sense of *bank* is dominated by the financial institution sense in the prototype representation). Thus, disambiguating the meaning of homonyms constitutes a valid falsification criterion for DSMs that posit abstraction at learning.

Although vector-space models of semantics offer sophisticated and inventive accounts of encoding, they offer weak theories of retrieval. For example, LSA presents a clever method for knowledge induction using singular value decomposition and dimension reduction. But it invokes a naïve theory for retrieval that ignores established wisdom from the study of human memory: Remembering is context dependent (e.g., Godden and Baddeley 1975), constructive (Bartlett 1932), and conditional on the interaction between how information is encoded and accessed (e.g., Morris et al. 1977; Tulving and Pearlstone 1966; Tulving and Thomson 1973). Tulving and Watkins 1973, p. 744) summarized the idea well:

> … retrieval always depends both on the availability of information in the memory store and on the accessibility of that information through appropriate retrieval cues, the latter being fragmentary knowledge the system possesses before retrieval about the material to be retrieved. While the exact mechanism remains to be specified, we find it helpful to think that the information contained in the retrieval cues somehow actively combines or interacts with the stored information to create the memory of a previously experienced event. Retrieval cues may vary greatly in their effectiveness, depending on the relation between the format of stored information and the encoding of the cues.

If we accept Tulving (1972) theorem, the vector-space approach to semantics presents only half of the solution for a complete account of semantics. How, then, can we implement the lessons from classical memory theory to develop a complete account of how language knowledge is stored and meanings are retrieved?

In the next section, we specify an instance-based model of memory that formalizes semantic abstraction as an episodic retrieval mechanism, rather than a learning mechanism. The model builds on storage and retrieval methods from Hintzman's (1984, 1986, 1988) MINERVA 2 theory of episodic memory and advances previous efforts (Dennis 2005; Johns and Jones 2015; Kwantes 2005) to demonstrate that semantic memory can be conceptualized as an artifact of retrieval from episodic memory rather than an encoding mechanism and separate store as outlined in Tulving's (1972) modular taxonomy of human memory.

Modeling Semantics as Retrieval from Episodic Memory

The instance theory of semantics (ITS) operates based on a combination of the encoding schemes of the BEAGLE model (Jones and Mewhort 2007) and the retrieval operations of the MINERVA 2 framework (Hintzman 1984, 1986, 1988). In this framework, every letter string (i.e., word or nonword) is represented by a unique n dimensional vector, w, where each dimension takes a randomly sampled value from a normal distribution with mean zero and variance $1/n$. Vectors constructed in this manner are orthonormal in expectation and are assumed to represent the physical form (orthography or phonology) of the word, as in BEAGLE.

Memory for an example of language is encoded as an instance context c_i, equal to the sum of the $j = 1 \ldots h$ word vectors in context i,

$$c_i = \sum_{j=1}^{j=h} w_{ij} \tag{1}$$

where h is the number of words in context i, w_j is word j in the context, and c_i is the sum of the words in context i. To illustrate, the context, "the dog bit the mailman" is stored as $w_{\text{dog}} + w_{\text{bit}} + w_{\text{mailman}}$ (consistent with standard practice, we excluded a list of stop words).

The memory for each context, c_i, is stored as a separate row in an m x n memory matrix, M, where rows correspond to memory traces (i.e., document contexts) and columns correspond to features,

$$M_i = c_i = \sum_{j=1}^{j=h} w_{ij} \tag{2}$$

Forgetting is treated as data loss and is implemented by deleting each feature in a memory with probability F. Thus, as F increases from 0 to 1, memory for documents in memory

degrades. Although the theory permits forgetting, we hold the parameter constant at 1 (i.e., no forgetting assumed).

To retrieve a word's meaning, a word vector is presented to memory as a probe and a corresponding semantic vector is retrieved that is called the *echo*. To compare word meanings, their echoes are compared. Words with matching echoes have identical meanings; words with nonmatching echoes match in meaning in proportion to the match.

Retrieving the echo is a two-step process. In step 1, the probe, p, activates all traces in memory, M, in parallel. Each trace's activation is computed as the cube of its cosine similarity to the probe,

$$a_i = \left(\frac{\sum_{j=1}^{j=n} p_j \times M_{ij}}{\sqrt{\sum_{j=1}^{j=n} p_j^2} \sqrt{\sum_{j=1}^{j=n} M_{ij}^2}} \right)^3 \quad (3)$$

where, a_i is the activation of trace i in memory, p_j is the value of feature j in the probe, M_{ij} is the value of feature j of trace i in memory, and n is the number of columns in memory. Activation ranges between − 1 and + 1. When the trace and probe are identical $a = 1$, when the trace and probe are orthogonal $a = 0$, and when the trace and probe are opposite $a = -1$.

Figure 1 shows the relationship between a trace's similarity to the probe and its activation. As shown, trace i's activation, a_i, is defined as a positively accelerated transformation of its cosine similarity to the probe. The nonlinearity of the activation function allows retrieval to be quite selective so that only the traces that are very similar to the probe are strongly activated and thus retrieved into the echo. The activation function is an important feature that makes the theory an instance theory and, as we will show, is critical for the theory to retrieve selectively and solve the disambiguation problem.

Second, a weighted sum of the traces is retrieved, where each trace is weighted by its corresponding activation,

$$e_j = \sum_{i=1}^{i=m} \sum_{j=1}^{j=n} a_i \times M_{ij} \quad (4)$$

where e_j is feature j in the echo, m is the number of traces in memory, a_i is the activation of trace i, and M_{ij} is the value of feature j in trace i in memory. The echo is the corresponding semantic representation retrieved for the probe.

Finally, the semantic resemblance, r, between two probes, p_1 and p_2, is computed as the cosine similarity between their corresponding echoes,

$$r(p_1, p_2) = \frac{\sum_{j=1}^{j=n} e_{1j} \times e_{2j}}{\sqrt{\sum_{j=1}^{j=n} e_{1j}^2} \sqrt{\sum_{j=1}^{j=n} e_{2j}^2}} \quad (5)$$

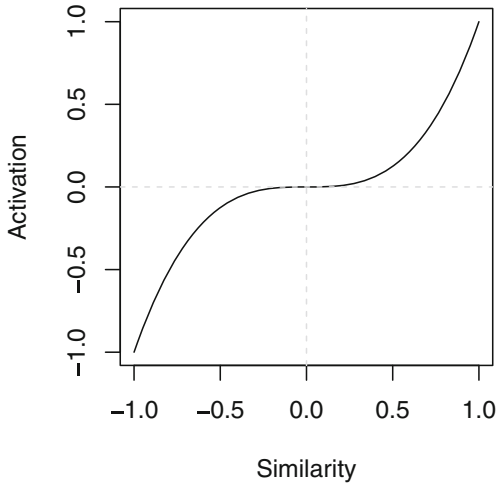

Fig. 1 Transformation of similarity to activation in the instance-based model

In summary, the theory assumes that humans record memory of their language experience, where each experience produces a trace that stores a sum of the words encountered. In contrast to most current accounts of semantics that apply a prototype abstraction algorithm to encode word meaning, ITS records a neutral record of word use in the corpus and develops a representation of word meaning on-the-fly by parallel and probe-driven retrieval. The model's on-the-fly derivation of semantics, conditional on the probe that is presented to memory, is the critical feature that distinguishes it from most other models of semantics (cf. Kwantes 2005) and that, we will show, allows it to explain and track context-specific retrieval of word meaning. We now turn to a demonstration of the theory using an engineered and very simple artificial language.

Experiment 1: Homonyms in an Artificial Language

If ITS learns word meanings, it should predict human judgments of word meaning. However, natural language is a complex structure rife with ambiguity and confounding. Thus, we begin with a contrived demonstration using a toy language.

Elman (1990) evaluated a simple recurrent network model of language against a toy artificial language (see Elman's 1990, Tables 3 and 4). His toy language included 13 different word classes, each represented by one or two words. For example, the word class NOUN_HUM was represented by the words *man* and *woman*, the word class VERB_EAT was represented by the word *eat*, and the word class NOUN_FOOD was represented by the words *cookie* and *sandwich*. The language also included 15 sentence templates used to construct language tokens (i.e., sentences). For example, the template NOUN_HUM, VERB_EAT, NOUN_FOOD can be used to

produce the following four sentences (a) "Man eat cookie", (b) "Man eat sandwich", (c) "Woman eat cookie", and (d) "Woman eat sandwich" by applying the following rewrite rules: (1) NOUN_HUM → {*man*, *woman*}, (2) VERB_EAT → {*eat*}, and (3) NOUN_FOOD → {*cookie*, *sandwich*}. To evaluate his SRN model of language, Elman trained the network on a corpus of sample sentences generated from this language.

We constructed an even simpler version of Elman's (1990) language to test our model. Our artificial language is presented in Table 1 and includes seven word classes, each represented by two words (e.g., NOUN_HUMAN is represented by two words *man* and *woman*). Critical for the analysis that follows, the word *break* was included in all three of the verb classes: VERB_VEHICLE → {*stop*, *break*},[1] VERB_DINNERWARE → {*smash*, *break*}, and VERB_NEWS → {*report*, *break*}. Thus, *break* is a homonym with three unrelated senses.

The use of *break* in all three verb classes was a deliberate component of our design and makes *break* an ambiguous word with three orthogonal meanings: In the vehicle sense, it is related to *stop*; in the dinnerware sense, it is related to *smash*; and in the news sense, it is related to *report*. We will use this feature of the language to test our model's ability (a) to appreciate a homonym's semantic ambiguity when presented in isolation and (b) to disambiguate the cued meaning of a homonym when presented in context (e.g., *break*/*story* → *report*).

The language includes three sentence frames. The first sentence frame produces sentences about *stopping*/*breaking* (i.e., braking) vehicles. The second sentence frame produces sentences about *smashing*/*breaking* dinnerware. The third sentence frame produces sentences about *reporting*/*breaking* news.

To generate a corpus, we sampled 20,000 sentences from the language. To generate a sentence, we (a) sampled one of the three sentence frames and (b) substituted words for the word classes in the sentence frame. For example, after sampling the sentence frame {NOUN_HUMAN, VERB_VEHICLE, NOUN_VEHICLE}, we sampled words *man*, *stop*, and *car* to the three respective word classes, thus producing the sentence "man stop car."

Applying the Model

Once a corpus had been generated, we applied our model in four steps. First, we generated a random vector of dimensionality 20,000 for each of the 12 words in the language (i.e., *man*, *woman*, *car*, *truck*, *plate*, *glass*, *story*, *news*, *stop*, *break*, *smash*, and *report*).[2] Second, we stored a representation of each sentence as a trace in memory; thus, memory was a 20,000 (contexts) by 20,000 (dimensions) matrix. Third, we retrieved an echo for each of the individual words. Finally, we computed the similarity between the echo retrieved for each word against the echo retrieved for each of the other 11 words.

The top panel in Fig. 2 shows the semantic relationships between all 12 words as a two-dimensional plot derived using MDS on the corresponding echoes (Shepard 1980). As shown, the monogamous words from the vehicle topic are clustered together (i.e., *stop*, *car*, *truck*), the monogamous words from the dinnerware topic are clustered together (i.e., *plate*, *glass*, *smash*), the monogamous words from the news topic are clustered together (i.e., *story*, *news*, *report*), and the promiscuous words (i.e., words that occurred in all three topics) are clustered together (i.e., *man*, *woman*, and *break*). Secondly, the different topic clusters are separated and distinct in space. Thirdly, the nouns within each cluster are closer to one another than they are to their corresponding monogamous verb (e.g., *car* and *truck* are closer to one another than they are to *stop*). Fourthly, the ambiguous word (i.e., *break*) is equidistant to the vehicle, dinnerware, and news clusters.

The results confirm that an instance-based approach to semantics can recover the structure of a small artificial language and that it can recognize semantic ambiguity. But, can the theory disambiguate the meaning of the ambiguous word (i.e., *break*) when it is presented in context (e.g., *break*/*glass* → *smash*)?

Using Context to Disambiguate Word Meaning

The corpus we generated establishes that the word *break* has three equally likely meanings. In the vehicle context, *break* (i.e., brake) is used synonymously with *stop*, as in stopping or braking a car. In the dinnerware context, *break* is used synonymously with *smash*, as in smashing or breaking a dinner plate. In the news context, *break* is used synonymously with *report*, as in reporting or breaking a story. The solution in the bottom panel of Fig. 2 shows that presenting *break* in isolation to ITS retrieves an echo that is equally similar to all three of its potential meanings.

In the next simulation, we present *break* in conjunction with a disambiguating word to determine if the echo retrieved by the joint probe (e.g., *break*/*car*) retrieves an echo that is more similar to the contextually cued meaning (i.e., *stop*) than

[1] Ignore the misspelling of *break* in the vehicular sense (i.e., brake). If the language is auditory, then the phonology of the break-brake homophone is identical, and so we use a single spelling (break) here so the word has an identical input to the model in either verb sense.

[2] We could have used a substantially smaller dimensionality for the word vectors, but very high dimensionality vectors allowed us to derive stable semantic representations later in the paper when we apply the theory to a large corpus of natural language.

Table 1 Artificial language

Categories of lexical items		
Categories	Examples	
NOUN_HUMAN	*man*, *woman*	
NOUN_VEHICLE	*car*, *truck*	
NOUN_DINNERWARE	*plate*, *glass*	
NOUN_NEWS	*story*, *news*	
VERB_VEHICLE	*stop*, ***break***	
VERB_DINNERWARE	*smash*, ***break***	
VERB_NEWS	*report*, ***break***	
Sentence frames		
NOUN_HUMAN	VERB_VEHICLE	NOUN_VEHICLE
NOUN_HUMAN	VERB_DINNERWARE	NOUN_DINNERWARE
NOUN_HUMAN	VERB_NEWS	NOUN_NEWS

to the competing contextually uncued meanings (i.e., *smash* or *report*).

To probe memory with a joint probe, we need to expand the activation function presented in Formula 3. To solve the problem, we borrowed Kwantes' (2005) reasoning (see also Estes' 1994 product rule) and computed a trace's activation as the product of the activations for all h individual words in the query:

$$a_i = \prod_{k=1}^{k=h} \left(\frac{\sum_{j=1}^{j=n} p_{kj} \times M_{ij}}{\sqrt{\sum_{j=1}^{j=n} p_{kj}^2} \sqrt{\sum_{j=1}^{j=n} M_{ij}^2}} \right)^3 \quad (6)$$

where a_i is the activation of trace i, p_{kj} is feature j of word k in the probe, M_{ij} is feature j of document i in memory, n is the dimensionality of a word representation, and h is the number of words in the probe.

The expanded activation function supports a selective activation of traces that retrieves traces that are similar to *all* words in the probe. Thus, even if one word in the probe is similar to the trace, pairing it with a word that is not similar to the trace will, by the product rule, cause the trace to be only weakly activated and thus weakly retrieved into the echo. Instances that contain most, or all, of the cue words will thus be preferentially activated. The mechanism should support contextual disambiguation. For example, presenting a joint probe *break*/*car* will weakly activate traces that are similar to only one of the two words, but will strongly activate traces that are similar to both words. The new activation function is perfectly consistent with the one presented in Eq. 3 (i.e., Eq. 6 simplifies to Eq. 3 when $h = 1$).

We evaluated ITS's contextual disambiguation of word meaning by retrieving an echo for *break* in conjunction with each of the six nouns (i.e., *car*, *truck*, *glass*, *plate*, *story*, and *news*) and, then, computing its similarity to the echoes retrieved with the three monogamous verbs (i.e., *stop*, *smash*, and *report*). If the model disambiguates the correct meaning of *break*, it should interpret *break*/*car* and *break*/*truck* as similar to *stop*, *break*/*plate* and *break*/*glass* as similar to *smash*, and *break*/*story* and *break*/*article* as similar to *report*.

Results are shown in the bottom panel of Fig. 2. As shown, ITS successfully retrieved the intended meaning of *break* in the context of a disambiguating partner. The first set of bars shows the similarity of *break* to each of its possible meanings (i.e., presenting *break* in isolation retrieves an echo that is equally similar to the echoes for *stop*, *smash*, and *report*). The null difference between the three senses establishes that the model knows all three meanings of the word *break* and that it knows them equally well.

The remaining bars in the bottom panel of Fig. 2 show the semantic similarity of *break* to each of its three possible senses, conditional on the partnered and disambiguating noun. As shown, when *break* was probed in conjunction with *car* or *truck*, the echo was more similar to *stop* than it was to either *smash* or *report*. When *break* was probed in conjunction with *plate* or *glass*, the echo was more similar to *smash* than it was to either *stop* or *report*. When *break* was probed in conjunction with *story* or *article*, the echo retrieved was more similar to *report* than *stop* or *smash*. Finally, and equally important to the positive disambiguation of the homonym *break*, the echoes were also *less similar* to their uncued meanings. For example, when *break* was probed in conjunction with *car* or *truck*, the echo was less similar to the competing alternative meanings (i.e., *smash* and *report*). The same pattern held for the other two senses of *break*.

In conclusion, the simulations in Fig. 2 show that ITS understands the overall ambiguous meaning of *break* when presented in isolation and the particular meanings of *break* when presented in conjunction with a disambiguating noun.

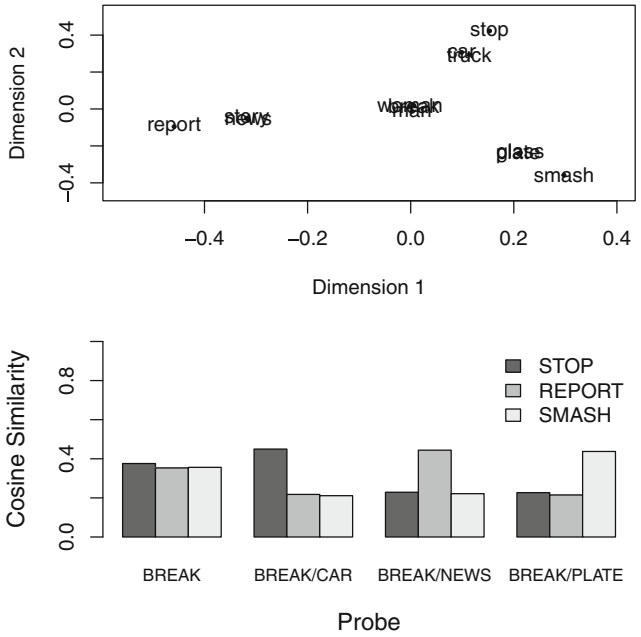

Fig. 2 Results from ITS with the artificial language, where the ambiguous word (i.e., *break*) occurs equally often in all three senses (i.e., vehicle, news, and dinnerware, all $p = 1/3$). The top panel shows the semantic space for all words in the language. The bottom panel shows ITS's ability to disambiguate the meaning of *break* depending on the context in which it is presented

Prototype Accounts

For comparison, and to assess the criticisms of Griffiths et al. (2005, 2007) of prototype accounts of distributional semantics, we conducted corresponding simulations using LSA (i.e., a first-generation prototype model of distributional semantics) and BEAGLE (i.e., a modern prototype model of distributional semantics).

In the simulations with LSA (Landauer and Dumais 1997), we (a) derived the word-by-context matrix from the corpus, (b) weighted the matrix by the standard entropy calculation, (c) derived a solution by dimension reduction, (d) computed the cosine similarity between words in each of the reduced spaces, and (e) tested disambiguation of the word *break* in conjunction with each of the six disambiguating nouns (i.e., *car*, *truck*, *plate*, *glass*, *story*, and *news*).

In the simulations with BEAGLE (Jones and Mewhort 2007), we constructed an environment vector, e_i, for each of the i words in the corpus (i.e., a vector of dimensionality $n = 2000$ where each dimension takes a value randomly sampled from a normal distribution with mean 0 and variance $1/n$) and derived the semantic vector for each of the $i = 1 \ldots 12$ words in the corpus by,

$$m_i = \sum_{j=1}^{j=c} \sum_{k=1}^{k=h} e_{jk} \tag{7}$$

for every sentence context that includes word i, where m_i is the semantic representation of word i, e_{jk} is word k in sentence j, s is the number of sentences in the corpus, and h is the number of words in a sentence. After the word representations had been constructed, we used the representations (i.e., all i memory vectors) to test for disambiguation of the word *break* in isolation and in conjunction with each of the six disambiguating nouns (i.e., *car*, *truck*, *plate*, *glass*, *story*, and *news*).

Because neither LSA nor BEAGLE store traces, we could not retrieve the semantic representation by presenting a probe and retrieving a representation using the joint activation function. Thus, we followed tradition and formed a joint probe by averaging the relevant word representations (e.g., *break*/*car*- = [*break* + *car*]/2) and computed the cosine similarity of this centroid representation to the vectors for the three verb senses (i.e., *stop*, *smash*, and *report*).

Results with both LSA and BEAGLE are presented in Figs. 3 and 4, respectively. The results are presented in the same format as the results for MINERVA to ease a visual comparison of the results over the three models.

The MDS solutions in Figs. 3 and 4 show that LSA and BEAGLE arrive at very similar solutions as ITS for individual words: Both recognize the semantic similarity of words within each of the three topic categories and recognize the difference in semantic similarity between the three topic categories. Naturally, this is expected.

However, in contradiction to conventional wisdom that prototype models of semantics should fail to disambiguate the contextually cued meaning of a homonym, the bottom rows of Figs. 3 and 4 show that LSA and BEAGLE succeed at disambiguating the cued meaning of a homonym. That is, in contradiction to accepted wisdom, they do understand the different senses of the homonym *break*, differently and appropriately, depending on the context in which *break* is presented.

The success of both LSA and BEAGLE at disambiguating the contextually appropriate meaning of *break* is surprising. However, the next simulation will show that the theories do in fact fail, as Griffiths et al. (2005, 2007) argued, under more realistic conditions.

Analysis of Disambiguation When the Homonym Has a Dominant Sense

It is well known that vector-based models of semantics have trouble disambiguating the meaning of homonyms. However, ITS, LSA, and BEAGLE all succeeded in our simple test. Why? One sensible criticism is that our toy language differs from natural language in several important ways (e.g., number of word classes, number of words, breadth of variation, difference in complexity, and so on). Consequently, the success might be illusory. In the next simulation, we reassessed ITS, LSA, and BEAGLE using a different corpus from the same

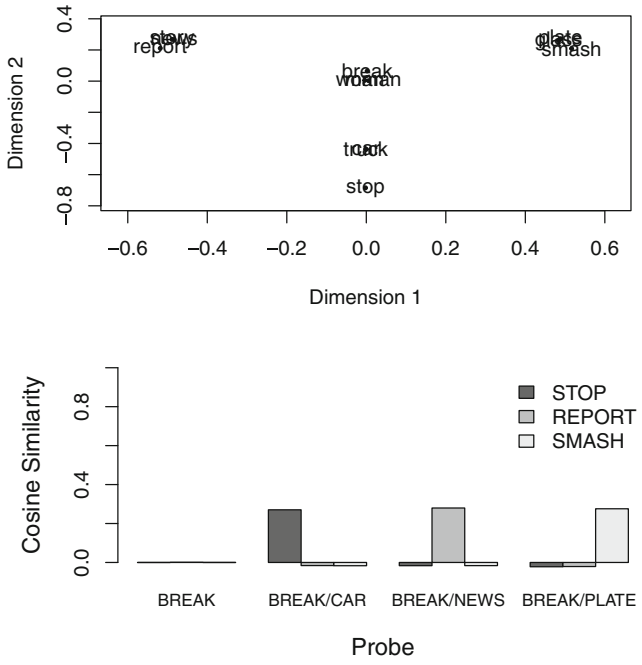

Fig. 3 Results from LSA with the artificial language, where the ambiguous word (i.e., *break*) occurs equally often in all three senses (i.e., vehicle, news, and dinnerware, all $p = 1/3$). The top row shows the semantic space for all words in the language. The bottom row shows LSA's ability to disambiguate the meaning of *break* depending on the context in which it is presented

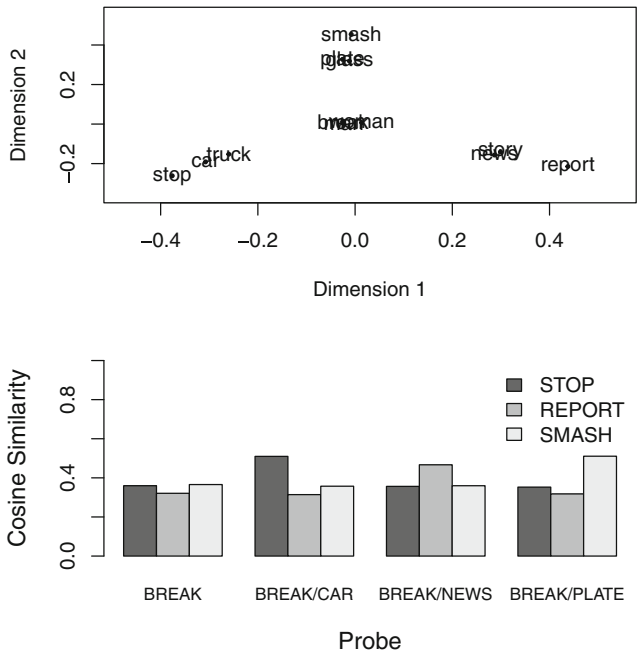

Fig. 4 Results from BEAGLE with the artificial language, where the ambiguous word (i.e., *break*) occurs equally often in all three senses (i.e., vehicle, news, and dinnerware, all $p = 1/3$). The top row shows the semantic space for all words in the language. The bottom row shows BEAGLE's ability to disambiguate the meaning of *break* depending on the context in which it is presented

language that was constructed so that *break* had a dominant sense.

To give *break* a dominant sense, we constructed a new corpus where sentences from the vehicle topic appeared more frequently than sentences from the dinnerware and news topics: p(NOUN_HUMAN, VERB_VEHICLE, NOUN_VEHICLE) = 4/6, whereas p(NOUN_HUMAN, VERB_DINNERWARE, NOUN_DINNERWARE) = p(NOUN_HUMAN, VERB_NEWS, NOUN_NEWS) = 1/6. Although our manipulation of the corpus is modest (i.e., the dominant sense could be engineered to be even more dominant), our simulations will show that changing the distribution of context-specific word use has a measurable and, as we will demonstrate, diagnostic influence on model performance. To foreshadow the results, our simulations will substantiate the common criticism of vector-based distributional approaches to semantics. However, our simulations will also show that our instance-based approach to semantics elegantly handles the problem, a selective activation of memory instances that are consistent with a cued subordinate sense.

The simulation we conducted was identical to the one already presented. However, the corpus was constructed so that sentences from the vehicle context were more frequent than sentences from the dinnerware and news contexts. The results from the simulations are presented in Figs. 5, 6, and 7 for simulations with ITS, LSA, and BEAGLE, respectively.

As shown in Fig. 5, the behavior of ITS was affected by the manipulation, but in sensible ways. Firstly, the monogamous words within each topic cluster together as before. Secondly, the promiscuous words (i.e., words that appeared in all three topic contexts) remain clustered, but consistent with the change in topic, base rates are closer to the monogamous words from the vehicle topic (i.e., the context in which they appeared more frequently) than the news and dinnerware topics. Finally, and most critically, the manipulation had no noticeable effect on the model's ability to disambiguate the meaning of the homonym *break* when *break* was presented in the context of a disambiguating noun. As shown in Figs. 6 and 7, the same is not true of the prototype models.

As shown in Fig. 6, LSA's behavior was strongly affected by the manipulation. Firstly, LSA concludes that the words in the dinnerware and news categories are similar to one another (i.e., by virtue of their shared difference to words in the vehicle category). Secondly, presenting LSA with *break*/*car* or *break*/*truck* (i.e., the dominant sense) recovers the cued sense of *break* (i.e., *stop*); however, presenting LSA with *break*/*plate*, *break*/*glass*, *break*/*story*, and *break*/*news* (i.e., the subordinate sense) does not—eliciting both *smash* or *report* as equally possible senses.

As shown in Fig. 7, BEAGLE produces a pattern of single word results that is very similar to the one produced by ITS. However, as shown in the test for disambiguation of *break*

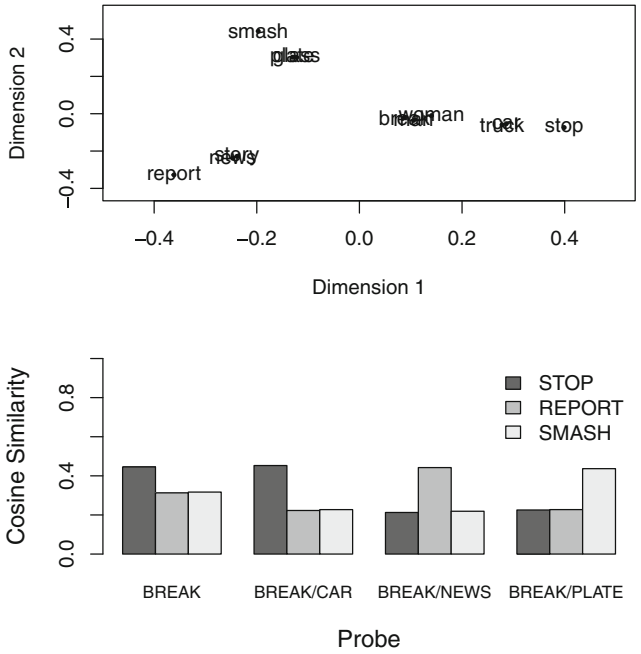

Fig. 5 Results from ITS with the artificial language, where the ambiguous word (i.e., *break*) occurs in one sense (i.e., p[*break*|vehicle] = 4/6) more than the others (i.e., p[*break*|news] = p[*break*|dinnerware] = 1/6). The top row shows the semantic space for all words in the language. The bottom row shows ITS's success at disambiguating the meaning of *break* depending on the context in which it is presented

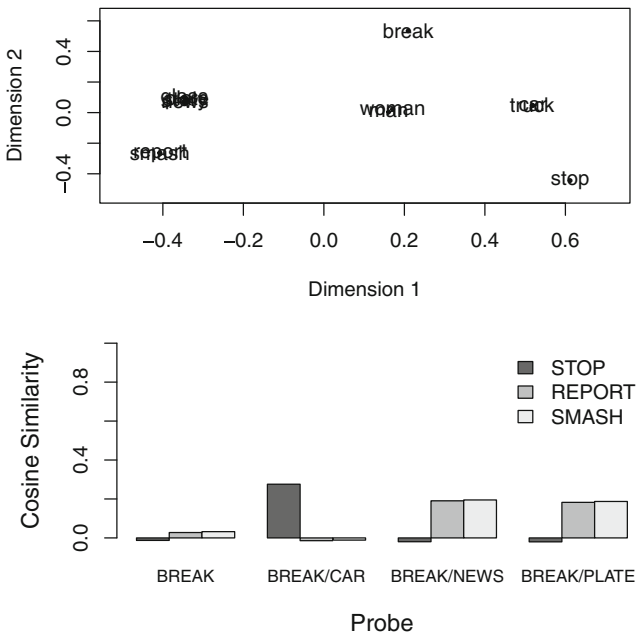

Fig. 6 Results from LSA with the artificial language, where the ambiguous word (i.e., *break*) occurs in one sense (i.e., p[*break*|vehicle] = 4/6) more than the others (i.e., p[*break*|news] = p[*break*|dinnerware] = 1/6). The top row shows the semantic space for all words in the language. The bottom row shows LSA's failure to disambiguate the meaning of *break* depending on the context in which it is presented

when presented in context, BEAGLE failed to understand the meaning of *break* in context—thinking instead that the meaning of *break* was consistent with its dominant meaning in all cases (i.e., *stop*).

The demonstration confirms the criticism of Griffiths et al. (2005, 2007) of the distributional prototype models. Unless semantically ambiguous words appear equally often in all senses, prototype models (i.e., LSA and BEAGLE) fail to understand the contextually valid meaning of that word. To the extent that LSA and BEAGLE fail to comprehend subordinate senses of homonyms, the common criticism is accurate: the prototype approach to distributional semantics offers a poor account of human behavior.

But, in contrast to the prototype models, ITS handles the disambiguation problem elegantly. Firstly, ITS recognizes the dominant sense of *break*: when *break* is presented in isolation, the echo retrieved was closer to *stop* than *smash* or *report* (i.e., the dominant sense). Secondly, even though ITS appreciates that *break* has a dominant sense, presenting *break* with a disambiguating noun (e.g., *break*/*plate*) allows it to retrieve the appropriately cued subordinate sense (i.e., *smash*). Finally, as in the previous simulations, presenting *break*/*car* or *break*/*truck* retrieves an echo more similar to the echo for *stop* than either *smash* or *report*, presenting *break*/*plate* or *break*/*glass* retrieves an echo more similar to *smash* than it does for either *stop* or *report*, and presenting *break*/*story* or *break*/*news* retrieves an echo more similar to the echo for *report* than it does for either *stop* or *smash*.

In summary, the criticism of Griffiths et al. (2005, 2007) of distributional models of semantics applies as charged to the prototype theories and, by extension, to similar models like HAL and Word2Vec. However, an instance-based approach to semantics is immune to this shortcoming because the architecture affords nonlinear activation of instances, which can reinstate the less common linguistic memories of the word. ITS only retrieves the traces in which both of the words in the joint probe occur. With the behavior of the instance-based theory now articulated in a controlled and contrived demonstration, we turn to an analysis of semantics derived from a natural language corpus.

Experiment 2: Natural Language Simulations

The simulations presented so far provide a good picture of our instance-based model of semantics and how it disambiguates the meaning of a homonym presented in context. However, solving a toy problem does not guarantee a solution to the problem at scale (e.g., Feldman-Stewart and Mewhort 1994).

To examine the theory at scale, we followed tradition and stored a record of language experience represented by the Touchstone Applied Science Associates (TASA) corpus,

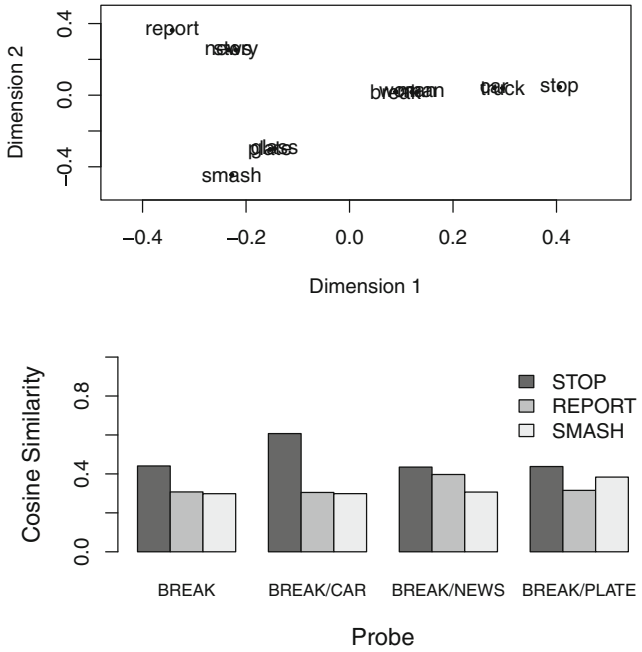

Fig. 7 Results from BEAGLE with the artificial language, where the ambiguous word (i.e., *break*) occurs in one sense (i.e., p[*break*|vehicle] = 4/6) more than the others (i.e., p[*break*|news] = p[*break*|dinnerware] = 1/6). The top row shows the semantic space for all words in the language. The bottom row shows BEAGLE's failure to disambiguate the meaning of *break* depending on the context in which it is presented

where each document in the corpus was encoded as a trace in memory (i.e., as a vector equal to a sum of the word vectors that correspond to the words in the document).

Next, we presented single and joint probes to memory and retrieved their echoes (i.e., their semantic representations). Finally, we computed the similarity between the echoes that were retrieved. We note that the model is applied exactly as it was in the previous simulations with the artificial language corpus and that only the corpus differs between the simulations that follow and the simulations already presented. Thus, an understanding of the model gained from the previous examples using the artificial language can be transferred wholesale for understanding the more expansive yet methodologically and computationally equivalent analysis of natural language semantics that follows.

Taxonomic Structure

A benchmark for semantic theories is that they can organize words into coherent taxonomic categories. For example, a competent theory of semantics should recognize that items from the category of *animals* are more similar to one another than they are to items from the category of *vehicles*. To evaluate our theory against the criterion of taxonomic organization, we stored a record of language experience from the TASA corpus and then retrieved echoes for words from defined taxonomic categories. For breadth, we derived solutions for words taken from a previous demonstration with BEAGLE (Jones and Mewhort 2007) and a previous demonstration with HAL (Lund and Burgess 1996).

Results for different tests are presented in Fig. 8 as two-dimensional MDS solutions. Plots in the left column of Fig. 4 present solutions for words from the three taxonomic categories examined in Jones and Mewhort (2007, Fig. 3): financial, science, and sports. The plots in the right column of Fig. 8 present solutions for words from three taxonomic categories examined in Lund and Burgess (1996, Fig. 2): body parts, countries, and animals. The complete categorized word lists are presented in Table 2.

As shown in the top row of Fig. 8, ITS does an excellent job of grouping words from the same categories while distinguishing words in opposing categories. To confirm the visual impression, we computed the intra- and intercategory similarities between words. For the Jones and Mewhort (2007) graph, the mean intracategory item-to-item cosine similarity ($M = 0.27$, SD = 0.11) was, by a conservative estimate, 1.82 standard deviations greater than the mean intercategory item-to-item similarity ($M = 0.07$, SD = 0.04). The same is true for the Lund and Burgess (1996) graph: the mean intracategory item-to-item cosine similarity ($M = 0.18$, SD = 0.10) was, by a conservative estimate, a still strong 1 standard deviations greater than the mean intercategory item-to-item similarity ($M = 0.08$, SD = 0.05).

For direct comparison, we conducted corresponding analyses using LSA and BEAGLE. Simulations with LSA used the vectors developed from the TASA corpus by Günther et al. (2015). Simulations with BEAGLE used vectors that we derived from the TASA corpus.

Results with LSA are presented in the middle row of Fig. 8. As shown, LSA positions items in the same category as more similar to one another than items in different categories. For the Jones and Mewhort (2007) set, the mean intracategory item-to-item cosine similarity ($M = 0.42$, SD = 0.28) was 1.50 standard deviations greater than the mean intercategory item-to-item similarity ($M = 0.00$, SD = 0.03). For the Lund and Burgess (1996) set, the mean intracategory item-to-item cosine similarity ($M = 0.14$, SD = 0.19) was 0.68 standard deviations greater than the mean intercategory item-to-item similarity ($M = 0.01$, SD = 0.05).

Results with BEAGLE are presented in the bottom row of Fig. 8. As shown, BEAGLE also rates words that belong to the same category as more similar to one another than to items in the opposing categories. For the Jones and Mewhort (2007) set, the mean intracategory item-to-item cosine similarity ($M = 0.48$, SD = 0.13) was 2.54 standard deviations greater than the mean intercategory item-to-item cosine similarity ($M = 0.15$, SD = 0.09). For the Lund and Burgess (1996) set, the mean intracategory

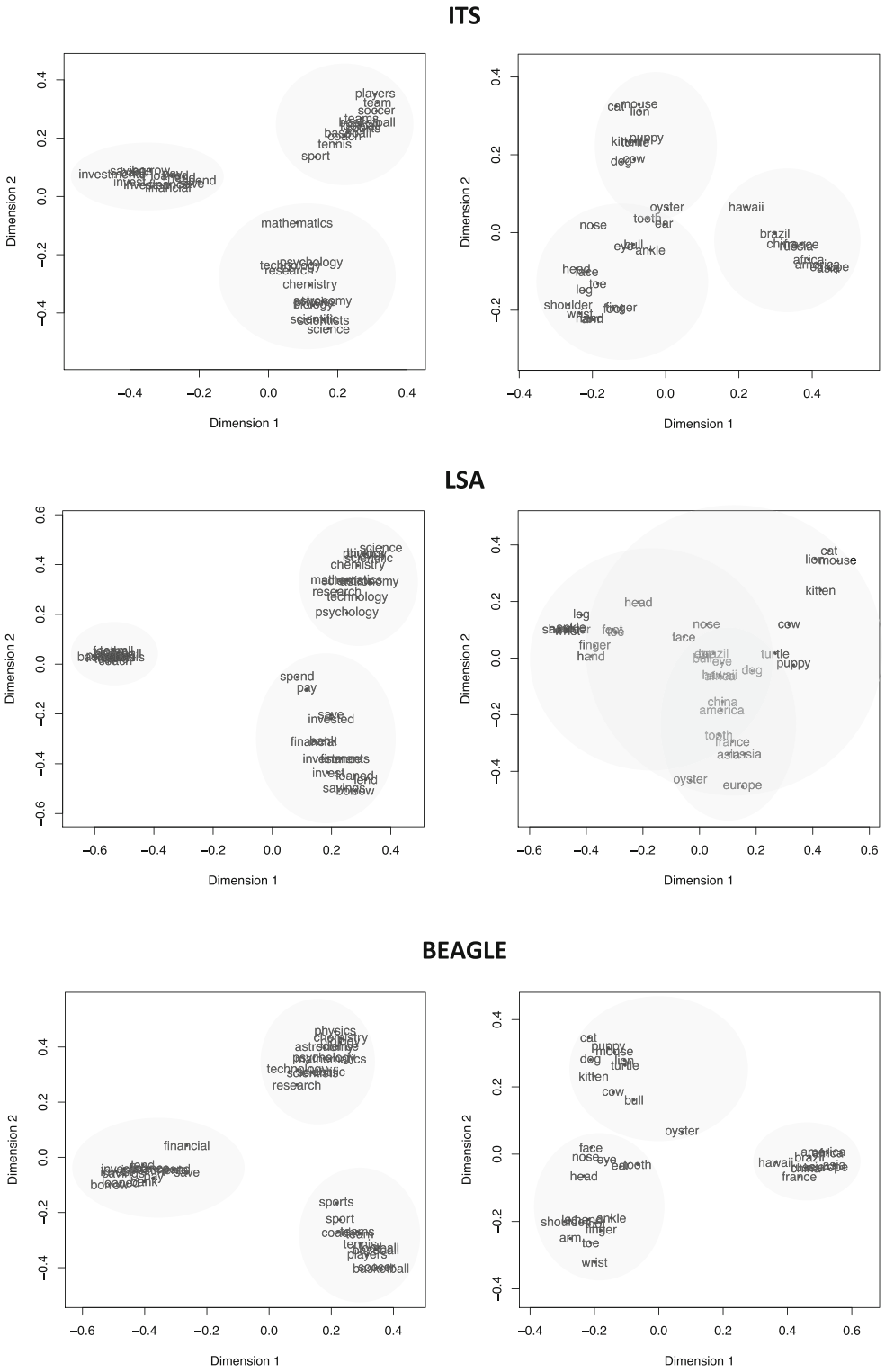

Fig. 8 Semantic spaces and taxonomic structure. The top panel presents results with ITS; the middle panel presents results with LSA; the bottom panel presents results with BEAGLE. Results on the left show results for words from Jones and Mewhort (2007). Results on the right show results for words from Lund and Burgess (1996)

item-to-item cosine similarity ($M = 0.40$, SD = 0.17) was 1.36 standard deviations greater than the mean intercategory item-to-item cosine similarity ($M = 0.17$, SD = 0.10).

Taken together, all three models do a good job of identifying which words belong to the same category and which do not. Critical to our analysis, the results serve as proof of concept that ITS, an instance-based model of distributional semantics, can perform taxonomic classification and that it performs within the same range of accuracy as LSA and BEAGLE, the standards for prototype models of distributional semantics.

Disambiguating Meaning in Context

The analysis of taxonomic structure provides a demonstration that ITS can group words that have related meanings. However, the demonstration does not provide evidence that the theory can disambiguate the meanings of homonyms conditional on context. To test the ability of ITS to understand the contextually appropriate meaning of a homonym, we conducted a simulation for disambiguation of homonyms using materials from a classic experiment reported by Schvaneveldt et al. (1976).

In their experiment, participants performed lexical decision. On each trial, three successive letter-strings were presented (e.g., *save-bank-money* or *save-bank-boat*) and the subject's task was to identify each string as a word or nonword as quickly and accurately as possible. On *cued trials*, the first two strings cued the appropriate meaning of the third string. For example, *save*/*bank* cued *money* and *river*/*bank* cued *boat*. On *miscued trials*, the first two words miscued the appropriate meaning. For example, *save*/*bank* miscued *boat* and *river*/*bank* miscued *money*. The critical result (or at least the one relevant here) was that participants were faster to identify the third word on cued than miscued trials.

To evaluate ITS, we conducted a simulation using the materials of Schvaneveldt et al. (1976) materials (see their Table 2, p. 248). On each trial, (a) an echo was retrieved for the joint probe composed of the first and second words (e.g., *save*/*bank*), (b) an echo was retrieved for the third word (e.g., *money*), and (c) the two echoes were compared. The cosine similarity of the two echoes indexed how well words 1 and 2 activated word 3 in the series.

We conducted a full set of comparisons to match the original experiment that included all 144 of the possible cued tests and 288 of the possible miscued tests. Finally, we computed a mean and variance for the cosines over all of the cued tests and the same for all of the miscued tests. We reasoned that if ITS disambiguates the meaning of a homonym presented in context, then the similarity between the echoes for joint primes and their targets will be reliably greater on cued than miscued trials. For example, cos(*save*/*bank*, *money*) > cos(*river*/*bank*, *money*).

The top leftmost panel of Fig. 9 shows the mean cosine similarity between the two-word cues and their targets, averaged over all 144 cued and 288 miscued trials.

There are three key differences to note. Firstly, the cosine similarity of the echo retrieved by the joint word primes was greater on cued than miscued trials. Secondly, the difference was statistically significant, greater than two standard errors. Thirdly, the standard error on miscued trials was greater than the standard error on cued trials—a difference that is consistent with the results of Schvaneveldt et al. (1976) results. In summary, an instance-based approach to semantics disambiguates homonyms in a manner consistent with data from human behavior.

Table 2 Words from Jones and Mewhort (2007) and from Lund and Burgess (1996) that were used in our tests of taxonomic organization, along with the categories to which they belong

Jones and Mewhort (2007)		Lund and Burgess (1996)	
Category	Word	Category	Word
Finance	Financial	Body part	Ankle
	Savings		Leg
	Finance		Shoulder
	Pay		Toe
	Invested		Finger
	Loaned		Wrist
	Borrow		Nose
	Lend		Ear
	Invest		Eye
	Investments		Hand
	Bank		Face
	Spend		Arm
	Save		Head
Science	Astronomy		Foot
	Physics	Countries	China
	Chemistry		Asia
	Psychology		France
	Biology		Russia
	Scientific		Europe
	Mathematics		Brazil
	Technology		Africa
	Scientists		America
	Science		Hawaii
	Research	Animals	Oyster
Sports	Sports		Puppy
	Team		Kitten
	Teams		Mouse
	Football		Dog
	Coach		Cow
	Sport		Cat
	Players		Lion
	Baseball		Bull
	Soccer		Turtle
	Tennis		Tooth
	Basketball		

Note. Lund and Burgess (1996) included *tooth* in the animal category rather than body part category, even though it could conceivably belong in either for different semantic reasons

For the sake of comparison, and to evaluate the criticism of Griffiths et al. (2005, 2007) of prototype accounts of distributional semantics, we repeated the simulation with LSA using the vectors developed by Günther et al. (2015) and the vectors that we derived using BEAGLE. As in our earlier simulations with the artificial language, we used the centroid method to

represent a joint probe: comparing the average of words 1 and 2 to the representation for word 3 in a series.

The results for LSA are shown in the middle leftmost panel in Fig. 9 and the results with BEAGLE are shown in the bottom leftmost panel in Fig. 9. As shown, LSA and BEAGLE both disambiguate the meaning of word 3 in a series when primed with the average of words 1 and 2: The mean cosine is greater on cued than miscued trials.

At first blush, our results suggest that the claim that the prototype models cannot disambiguate homonyms is false. However, the variances suggest that the mean cosine similarities from our analysis using LSA and BEAGLE might hide a more nuanced pattern of results. To explore the issue, we examined the performance for the individual items in the stimulus set of Schvaneveldt et al. (1976).

Armstrong et al. (2012) collected participant ratings on the semantic ambiguity of homonyms. Based on those data, they computed each homonym's dominance score—a measurement that indexes its semantic uncertainty in the range 0 to 1, where 0 indicates the word is perfectly ambiguous (i.e.,

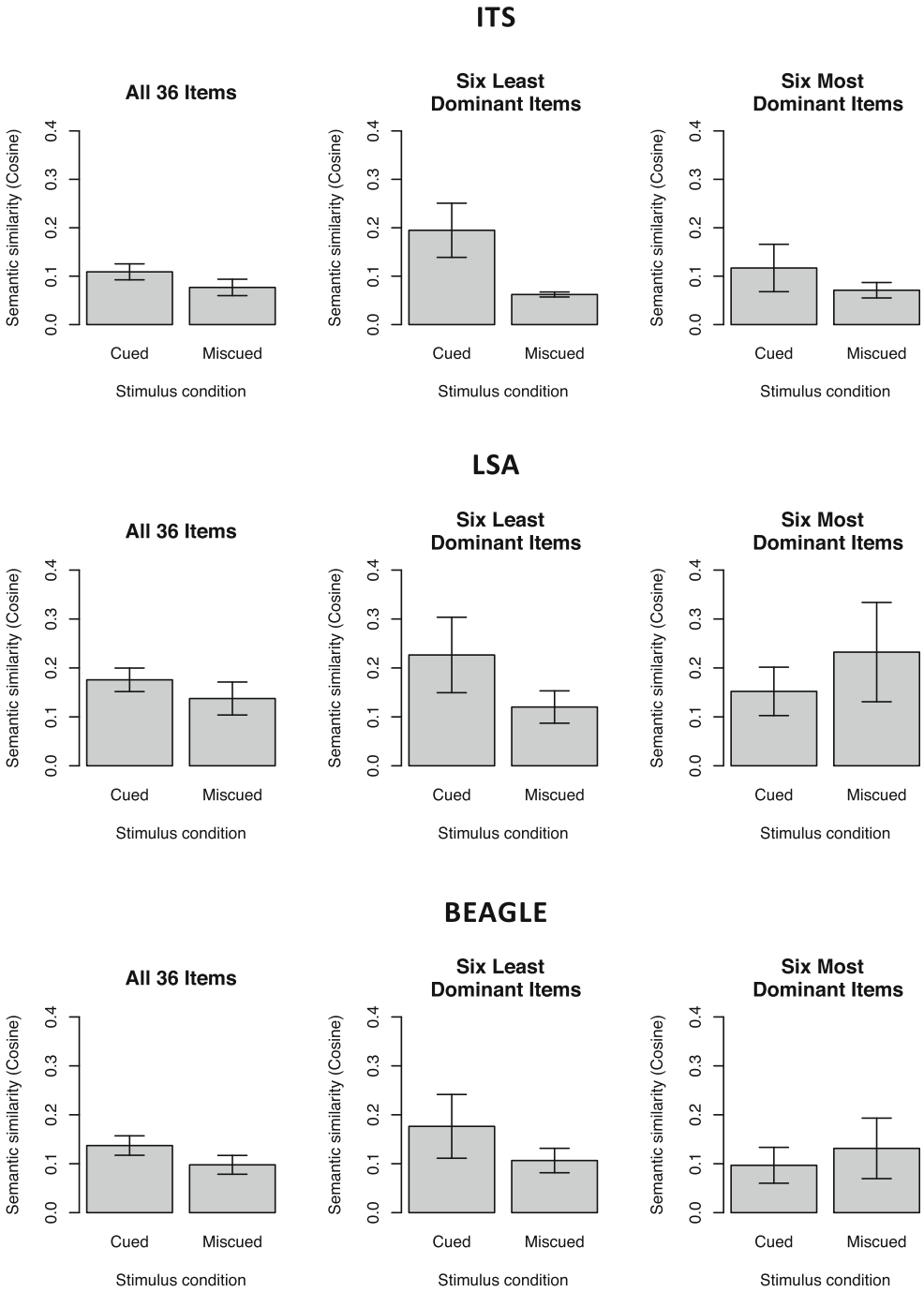

Fig. 9 Simulation of disambiguation with natural language from the Schvaneveldt et al. (1976, Table 2, p. 248) stimulus set. Results with ITS are shown in the top row. Results with LSA are shown in the middle row. Results with BEAGLE are shown in the bottom row. In all cases, semantic similarity is shown averaged over all 36 items from the stimulus set on the left, averaged over the six least dominant items in the center, and averaged over the six most dominant items on the right. Whiskers show the standard error of the mean computed over all comparisons in each condition

participants considered both potential meanings of a homonym to be equally likely) and 1 indicates the word is perfectly unambiguous (i.e., participants rate one meaning of a homonym to be absolute).

We cross-referenced the norms of Armstrong et al. (2012) with the stimulus list of Schvaneveldt et al. (1976) and copied the measurement of semantic dominance for the 27 of the 36 homonyms that appeared in both sets. A list of dominance scores for the 27 homonyms can be found in Table 3. As shown, participants rated some homonyms as having a strong dominant sense (e.g., *bank* is judged as having a dominant sense related to the place where money is kept versus the ridge of earth that separates water from land) and rated others as more ambiguous (e.g., the meaning of *bark* was rated as more uncertain in relation to its competing senses of the sound a dog makes and the material on the outside of a tree).

Table 3 Dominance scores from the norms of Armstrong et al. (2012) for 27 of the 36 homonyms in the stimulus list of Schvaneveldt et al. (1976)

Homonym	Dominance
Bail	0.46
Bank	0.68
Bark	0.35
Bowl	0.46
Box	0.43
Bridge	0.74
Calf	0.36
Fair	0.51
Fan	0.17
Fleet	0.66
Hide	0.60
Jam	0.28
Jar	0.70
Lying	0.49
Lock	0.66
Mint	0.52
Pen	0.65
Pit	0.61
Race	0.12
Ring	0.14
Sage	0.50
Spit	0.80
Stable	0.42
Stern	0.52
Tap	0.38
Tick	0.55
Tire	0.31

Note. The items *ball*, *date*, *fast*, *light*, *mine*, *mold*, *pick*, *punch*, and *riddle* from Schvaneveldt et al. (1976) were not included in the norms of Armstrong et al. (2012)

Using the empirical norms of Armstrong et al. (2012), we recalculated the difference in cosine similarities on cued versus miscued trials for the six homonyms in the stimulus list of Schvaneveldt et al. that had the smallest dominance scores (i.e., *bark*, *fan*, *jam*, *race*, *ring*, and *tire* that had uncertain interpretations; $M = 0.23$, SD = 0.10, range = 0.12 to 0.35) and the six that had the highest dominance scores (i.e., *bank*, *bridge*, *fleet*, *jar*, *lock*, and *spit*; $M = 0.71$, SD = 0.05, range = 0.66 to 0.80). An independent samples t test by items confirmed that the difference in the mean dominance score in the two sets (0.71–0.23 = 0.48) was statistically significant, $t(10) = 8.56$, $p < 0.0001$.

Model performances on the six least dominant and six most dominant homonyms are presented in the center and rightmost columns in Fig. 9, respectively, with results for ITS, LSA, and BEAGLE shown in the top, middle, and bottom panels, respectively. As shown, and consistent with the results from our analysis of the artificial language presented earlier (see Figs. 2, 3, 4, 5, 6, and 7), ITS succeeded at comprehending the meaning of a homonym in context (i.e., cued > miscued), whether the homonym was presented in its strong or weak sense. In contrast, both LSA and BEAGLE failed to comprehend the meaning of a homonym in context, unless it was presented in the context of its dominant sense (i.e., in the case of the strong dominance items, both models predict miscued > cued).

We conclude that the common criticism (see Griffiths et al. 2005, 2007) of LSA and, by corollary, related prototype models is valid when comprehending a homonym's subordinate sense. More importantly, our simulations show that the criticism *does not* apply to an instance-based approach to semantics that includes context-sensitive retrieval of meaning via selective and probe-driven retrieval of traces from memory.

General Discussion

Distributional models of semantics have been remarkably successful in cognitive science, both as tools and as theories of human learning and representation. But since their inception, the assumption has been that abstraction is a core goal of the learning system, storing a single economical representation that best captures the central tendency of the contexts in which a word occurs. This prototype assumption may have been implicitly guided by popular theoretical principles of cognitive economy (Rosch and Mervis 1975) and the Chomskian presumption that the job of the cognitive mechanism is to abstract the generative rules from language instances.

However, prototype models of semantics are at odds with other subfields of cognition, such as categorization and recognition, which have emphasized the superiority of instance-based and exemplar-based models over prototype models. As ITS demonstrates, a model that stores language instances

and applies a simple retrieval mechanism can produce on-the-fly semantic representations given a cue. Further, the model allows for nonlinear activation of instances, producing very different "abstracted" representations depending on a cue—something that is not possible if abstraction is applied at learning to derive a prototypical representation of each word (as with virtually all current and classic DSMs).

The notion that semantic abstraction may be better conceptualized as a retrieval mechanism rather than an encoding mechanism was originally posited by Kwantes (2005). Our ITS model builds on that advance using the architecture from Hintzman's (1984, 1986, 1988) MINERVA exemplar theory of human memory.

MINERVA was invented to explain how semantics can emerge during a probe-driven, selective, and parallel retrieval of instances. When the theory was developed, modeling of natural language semantics was limited by the available computational power and the relative paucity of research on natural language processing at scale—a topic that was only made tractable by the invention of the vector-based models of semantics (Landauer and Dumais 1997; Lund and Burgess 1996). Consequently, the argument was developed in relation to prototype learning of small artificial categories. But, our analysis is consistent with that initial examination and goal—just applied at the scale of natural language. Despite the difference in scale, we view our analysis to be, at its core, a straightforward restatement and extension of Hintzman's (1984, 1986, 1988) original thesis about category learning within the domain of language.

More generally, our demonstration shows that an established and classic theory for memory that has previously been applied to understand a suite of behaviors including (a) recognition memory (Hintzman 1984), (b) frequency judgment (Hintzman 1988), (c) cued recall (Hintzman 1986), (d) classification (Hintzman 1986), (e) function learning (Kwantes and Neal 2006), (f) judgment and decision (Dougherty et al. 1999; Thomas et al. 2008), (g) speech normalization (Goldinger 1998), (h) confidence/accuracy inversions in eyewitness identification (Clark 1997), (i) language processing (Rosch and Mervis 1975), (j) false memory (Arndt and Hirshman 1998), (k) memory dissociations in aging (Benjamin 2010), (l) implicit learning (Jamieson and Mewhort 2009a, 2010, 2011), (m) speeded choice (Jamieson and Mewhort 2009b), (n) associative learning (Jamieson et al. 2010b, 2012), (o) the production effect in recognition memory (Jamieson et al. 2016a), and (p) selective memory impairment in amnesia (Jamieson et al. 2010a; Curtis and Jamieson 2018) can also be used to understand semantics. The cross-lab and cross-domain effort represents the way that science ought to progress—by developing a general account of memory and its processes in a working computational system to produce a common explanation of behavior rather than a set of lab-specific and domain-specific theories for different behaviors (Newell 1973).

Although not fully reported here, ITS can reproduce any general semantic phenomena that have been used to support classic DSMs. But ITS is also able to explain subordinate senses of homonyms in context due to nonlinear activation of language instances, where prototype DSMs lose the distinction due to their aggregated representation. It is important to note that our demonstration with homonyms is not a minor flaw with prototype DSMs—it is a critical falsification criterion. More than half of all English words have multiple senses, and the distribution of sense frequencies is heavy-tail distributed. Humans understand the distinction among word senses easily, but prototype DSMs are heavily biased by the dominant sense in averaging, similar to the problem of classic XOR in categorization. Prototype DSMs lose the tail of the sense distribution, and the tail is where many of our word meanings in memory live.

Although prototype DSMs do not disambiguate word meaning in our simulations (Griffiths et al. 2005, 2007), ad hoc patches have been developed to address the problem. For example, Kintsch (2001) developed a predication operation to disambiguate word meaning in LSA and Cohen and Widdows (2016) developed a projection method to do the same (see also Erk and Padó 2008, and Reisinger and Mooney 2010, for multiprototype methods developed in the domain of computational linguistics). However, it is always true that a theoretical model can be developed to accommodate a behavior once the behavior is known and articulated. Our instance-based solution, on the other hand, does not require an ad hoc patch. Rather, disambiguation of word meaning, even when the meaning is subordinate, falls naturally out of first principles. Although Kintsch's and Cohen and Widdow's methods might present one account of cognitive processes in semantic disambiguation in natural language processing, our solution provides an alternative way to think about semantics and can serve as a motivated and articulate foil to analyze the problem more closely in future work.

ITS gives an elegant solution for computing word meaning that is grounded in classic principles of human memory. However, adopting the method comes at computational expense (Stone et al. 2011). In a prototype DSM, each new language experience is integrated into existing semantic knowledge: Therefore, storage demands do not increase and the derivation of word meaning is computed up front. In an instance-based DSM, each new language experience lays down a new trace in memory and word meaning is derived from that record on-the-fly; therefore, each new trace increases the demands on memory as well as the time to derive a word's meaning. Thus, prototype DSMs present a more computationally efficient way to measure semantics than instance-based DSMs. So, how do we balance the theoretical

insights gained from an instance-based approach to semantics against computational efficiency?

There are many differences between the brain and computational databases in how they represent and retrieve information. The search and abstraction processes used in human cognition need not be identical to efficient database search. Models of cognition have long assumed that memory exemplars can be activated in parallel, although the code we use to implement this in a model will usually use a loop routine. This is a distinct difference between the two disciplines: Looping through all exemplars is not an efficient method of, for example, word similarity matching, but it may well be the correct model of how humans do it. In our estimation, the practical constraints of current computational hardware should not be used as a scientifically valid reason to discard working models of human cognition such as the instance-based model of semantics presented here.

Although we have provided an instance-based DSM to encode word meaning from language experience, Storms et al. (2000)) have examined the distinction between an instance-based and prototype-based approach to natural language classification. In that work, they relied on the generalized context model (Nosofsky 1984, 1986) rather than the MINERVA2 framework. In some of their work, the evidence favored an instance-based conclusion (Smits et al. 2002; Voorspoels et al. 2008). In other work, their evidence favored an intermediate representation somewhere in between an instance and prototype representation (Verbeemen et al. 2007; Voorspoels et al. 2011). Taken together, their work suggests that it may be naive to pit the instance and prototype DSMs against one another as though they were mutually exclusive. Rather, it might be more productive to consider how representations from the instance-based and prototype DSMs differ, how they complement one another, and how the different levels of representation are coordinated in semantic cognition.

In some ways, it is tempting to see instance-based DSMs as "cheating." If the model stores all data, then it can compute an accurate semantic representation whenever one is needed. But the theoretical claim is profound in its proposal: We may not have semantic memory in the way that theorists have typically conceived of semantic memory. In place of the standard view, an instance-based approach to semantics proposes that a person's interpretation of the words they are reading is constructed on the fly, where meaning is an artifact of retrieving the visual patterns from episodic memory and that our phenomenology of meaning is continuously constructed as the interaction between stimuli, episodic memory, and the memory retrieval mechanism that mediates them (Kintsch and Mangalath 2011). But the instance-based approach should also put us at ease because they provide converging evidence that performance across multiple cognitive domains (e.g., categorization, recognition, semantics) might be explicable from the same core cognitive principles (Newell 1994; Surprenant and Neath 2013).

References

Armstrong, B. C., Tokowicz, N., & Plaut, D. C. (2012). eDom: norming software and relative meaning frequencies for 544 English homonyms. *Behavior Research Methods, 44*, 1015–1027.

Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: explanations from a global matching perspective. *Journal of Memory and Language, 39*, 371–391.

Aujla, H., Jamieson, R. K., & Cook, M. T. (2018). A psychologically inspired search engine. In *Lecture notes in computer science: high performance computing systems and applications*. Springer, Berlin (in press).

Bartlett, F. C. (1932). Remembering. Cambridge.

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., Ribeiro, S., Javitt, D. C., Copelli, M., & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*.

Benjamin, A. S. (2010). Representational explanations of "process" dissociations in recognition: the DRYAD theory of aging and memory judgments. *Psychological Review, 117*, 1055–1079.

Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169–211). Hillsdale: Erlbaum.

Brooks, L. R. (1987). Decentralized control of categorization: the role of prior processing episodes. In U. Neisser (Ed.), *Concepts and conceptual development: ecological and intellectual factors in categorization* (pp. 141–174). Cambridge: Cambridge University Press.

Clark, S. E. (1997). A familiarity-based account of confidence–accuracy inversions in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 232–238.

Cohen, T., & Widdows, D. (2016). Embedding probabilities in predication space with Hermitian holographic reduced representations. In H. Atmanspacher, T. Filk, & E. Pothos (Eds.), *Quantum interaction. QI 2015. Lecture notes in computer science* (Vol. 9535, pp. 245–257). Cham: Springer.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning & Verbal Behavior, 8*, 240–247.

Curtis, E. T., & Jamieson, R. K. (2018). Computational and empirical simulations of selective memory impairments: converging evidence for a single-system account of memory dissociations. Quarterly Journal of Experimental Psychology (in press).

Dennis, S. (2005). A memory-based theory of verbal cognition. *Cognitive Science, 29*, 145–193.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: a memory processes model for judgments of likelihood. *Psychological Review, 106*, 180–209.

Elman, J. L. (1990). Finding structure in time. *Cognitive science, 14*(2), 179–211.

Erk, K., & Padó, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 897–906). Association for Computational Linguistics.

Estes, W. K. (1994). Classification and cognition. Oxford University Press.

Feldman-Stewart, D., & Mewhort, D. J. K. (1994). Learning in small connectionist networks does not generalize to large networks. *Psychological Research, 56*, 99–103.

Firth, J. R. (1957). A synopsis of linguistic theory. *Studies in Linguistic Analysis*, 1930–1955.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*, 939–944.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In Advances in Neural Information Processing Systems (pp. 537–544).

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*, 211–244.

Godden, D., & Baddeley, A. (1975). Context dependent memory in two natural environments. *British Journal of Psychology, 66*, 325–331.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251–279.

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun—an R package for computations based on latent semantic analysis. *Behavior Research Methods, 47*, 930–944.

Hintzman, D. L. (1984). MINERVA-2: a simulation model of human memory. *Behavior Research Methods, Instruments & Computers, 16*, 96–101.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*, 411–428.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*, 528–551.

Jamieson, R. K., Crump, M. J. C., & Hannah, S. D. (2012). An instance theory of associative learning. *Learning & Behavior, 40*, 61–82.

Jamieson, R. K., Hannah, S. D., & Crump, M. J. C. (2010b). A memory-based account of retrospective revaluation. *Canadian Journal of Experimental Psychology, 64*, 153–164.

Jamieson, R. K., & Hauri, B. (2012). An exemplar model of performance in the artificial grammar task: holographic representation. *Canadian Journal of Experimental Psychology, 66*, 98–105.

Jamieson, R. K., Holmes, S., & Mewhort, D. J. K. (2010a). Global similarity predicts dissociation of classification and recognition: evidence questioning the implicit/explicit learning distinction in amnesia. *Journal of Experimental Psychology: Learning, Memory and Cognition, 36*, 1529–1535.

Jamieson, R. K., & Mewhort, D. J. K. (2005). The influence of grammatical, local, and organizational redundancy on implicit learning: an analysis using information theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 9–23.

Jamieson, R. K., & Mewhort, D. J. K. (2009a). Applying an exemplar model to the artificial-grammar task: inferring grammaticality from similarity. *Quarterly Journal of Experimental Psychology, 62*, 550–575.

Jamieson, R. K., & Mewhort, D. J. K. (2009b). Applying an exemplar model to the serial reaction time task: anticipating from experience. *Quarterly Journal of Experimental Psychology, 62*, 1757–1783.

Jamieson, R. K., & Mewhort, D. J. K. (2010). Applying an exemplar model to the artificial-grammar task: string-completion and performance for individual items. *Quarterly Journal of Experimental Psychology, 63*, 1014–1039.

Jamieson, R. K., & Mewhort, D. J. K. (2011). Grammaticality is inferred from global similarity: a reply to Kinder (2010). *Quarterly Journal of Experimental Psychology, 64*, 209–216.

Jamieson, R. K., Mewhort, D. J. K., & Hockley, W. E. (2016a). A computational account of the production effect: still playing twenty questions with nature. *Canadian Journal of Experimental Psychology, 70*, 154–164.

Jamieson, R. K., Nevzorova, U., Lee, G., & Mewhort, D. J. K. (2016b). Information theory and artificial grammar learning: inferring grammaticality from redundancy. *Psychological Research, 80*, 195–211.

Jamieson, R. K., Vokey, J. R., & Mewhort, D. J. K. (2017). Implicit learning is order dependent. *Psychological Research, 81*, 204–218.

Jones, M. N. (2017). *Big data in cognitive science*. United Kingdom: Psychology Press, Taylor & Francis.

Johns, B. T., & Jones, M. N. (2015). Generating structure from experience: A retrieval-based model of language processing. *Canadian Journal of Experimental Psychology , 69,* 233–251.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language, 55*, 534–552.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*, 1–37.

Johns, B. T., Taler, V., Pisoni, D. B., Farlow, M. R., Hake, A. M., Kareken, D. A., & Jones, M. N. (2013). Using cognitive models to investigate the temporal dynamics of semantic memory impairments in the development of Alzheimer's disease. In *Proceedings of the 12th international conference on cognitive modeling* (pp. 23–28).

Kintsch, W. (2001). Predication. *Cognitive Science, 25*, 173–202.

Kintsch, W., & Mangalath, P. (2011). The construction of meaning. *Topics in Cognitive Science, 3*(2), 346-370.

Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin & Review, 12*, 703–710.

Kwantes, P., & Neal, A. (2006). Why people underestimate y when extrapolating in linear functions. *Journal of Experimental Psychology: Learning Memory, and Cognition, 32*, 1019–1030.

Kwantes, P. J., Derbentseva, N., Lam, Q., Vartanian, O., & Marmurek, H. H. (2016). Assessing the Big Five personality traits with latent semantic analysis. *Personality and Individual Differences, 102*, 229–233.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*, 203–208.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems (pp. 3111–3119).

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Memory and Language*, 16(5), 519

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89*, 609–626.

Murdock, B. B. (1983). A distributed memory model for serial-order information. *Psychological Review, 90*, 316–338.

Murdock, B. B. (1995). Developing TODAM: three models for serial-order information. *Memory & Cognition, 23*, 631–645.

Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review, 104*, 839–862.

Newell, A. (1973). You can't play 20 questions with nature and win: projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing* (pp. 283–308). New York: Academic.

Newell, A. (1994). Unified theories of cognition. Harvard University Press.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 10*, 104–114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.

Recchia, G. L., Jones, M. N., Sahlgren, M., & Kanerva, P. (2010). Encoding sequential information in vector space models of semantics: comparing holographic reduced representation and random permutation. In S. Ohisson & R. Catrambone (Eds.), *Cognition in flux: Proceedings of the 32nd annual cognitive science society* (pp. 865–870). Austin: Cognitive Science Society.

Reisinger, J., & Mooney, R. J. (2010, June). Multi-prototype vector-space models of word meaning. In *Human language technologies: the 2010 annual conference of the North American Chapter of the Association for Computational Linguistics* (pp. 109–117). Association for Computational Linguistics.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology, 7*(4), 573–605.

Rubin, T. N., Koyejo, O., Gorgolewski, K. J., Jones, M. N., Poldrack, R. A., & Yarkoni, T. (2016a). Decoding brain activity using a large-scale probabilistic functional-anatomical atlas of human cognition. *bioRxiv*, 059618.

Rubin, T., Koyejo, O., Jones, M. N., & Yarkoni, T. (2016b). Generalized correspondence-LDA models (GC-LDA) for identifying functional regions in the brain. *Advances in Neural Information Processing Systems*.

Schvaneveldt, R. W., Meyer, D. E., & Becker, C. A. (1976). Lexical ambiguity, semantic context, and visual word recognition. Human perception and performance. *Journal of experimental psychology 2*(2), 243.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science, 210*, 390–398.

Simon, H. A. (1969). *The sciences of the artificial*. Cambridge: MIT Press.

Smits, T., Storms, G., Rosseel, Y., & De Boeck, P. (2002). Fruits and vegetables categorized: an application of the generalized context model. *Psychonomic Bulletin and Review, 9*, 836–844.

Stanton, R. D., Nosofsky, R. M., & Zaki, S. R. (2002). Comparisons between exemplar similarity and mixed prototype models using a linearly separable category structure. *Memory & Cognition, 30*, 934–944.

Stone, B., Dennis, S., & Kwantes, P. J. (2011). Comparing methods for single paragraph similarity analysis. *Topics in Cognitive Science, 3*, 92–122.

Storms, G., De Boeck, P., & Ruts, W. (2000). Prototype and exemplar based information in natural language categories. *Journal of Memory and Language, 42*, 51–73.

Surprenant, A. M., & Neath, I. (2013). Principles of memory. Psychology Press.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review, 115*, 155–185.

Tulving, E. (1972). Episodic and semantic memory. *Organization of Memory, 1*, 381–403.

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning & Verbal Behavior, 5*, 381–391.

Tulving, E., & Thomson, D. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*, 352–373.

Tulving, E., & Watkins, M. J. (1973). Continuity between recall and recognition. *The American Journal of Psychology*, 739–748.

Verbeemen, T., Vanpaemel, W., Pattyn, S., Storms, G., & Verguts, T. (2007). Beyond exemplars and prototypes as memory representations of natural concepts: a clustering approach. *Journal of Memory and Language, 56*, 537–554.

Voorspoels, W., Vanpaemel, W., & Storms, G. (2008). Exemplars and prototypes in natural language concepts: a typicality-based evaluation. *Psychonomic Bulletin and Review, 15*, 630–637.

Voorspoels, W., Vanpaemel, W., & Storms, G. (2011). A formal ideal-based account of typicality. *Psychonomic Bulletin and Review, 18*, 1006–1014.